ВОПРОСЫ АЛГОРИТМИЧЕСКОГО АНАЛИЗА СТРУКТУРНОЙ ИНФОРМАЦИИ
(Вычислительные системы)
1987 год Выпуск 119

УДК 541.I

ПРЕДСТАВЛЕНИЕ ХИМИЧЕСКИХ СОЕДИНЕНИЙ КАК РАСПРЕДЕЛЕНИЙ ФИЗИЧЕСКИХ СВОЙСТВ НА МОЛЕКУЛЯРНЫХ СТРУКТУРАХ

В.М.Зацепин

Введение

Традиционная структурная формула химического соединения с математической точки зрения представляет собой мультиграф с помеченными вершинами. Вершинам соответствуют атомы или структурные фрагменты, ребрам - связи.

В расширенном описании с вершинами и ребрами молекулярного графа могут быть связаны наборы меток, несущих содержательную (физическую) информацию, например, атомный номер, валентность, ван-дер-ваальсов объем, локальный заряд, инкременты молекулярной рефракции, логарифма распределения в системе "органическая фаза-вода", порядки связей и др. Фактически такого рода представление о молекуле и формируется, возможно в неявной форме, у химика при рассмотрении структурной формулы.

Таким образом, представление химических соединений как распределений физических свойств на молекулярных структурах кажется достаточно естественным. В данной работе рассматриваются некоторые вопросы их анализа. Молекулярная структура может быть представлена в виде помеченного графа либо графа в пространстве путем задания декартовых координат атомов или в виде матрицы межатомных расстояний.

Анализ молекулярных структур в химической информатике основан в настоящее время на преобразовании исходной структуры в набор булевых и/или количественных дескрипторов, характеризующих наличие/отсутствие определенных структурных фрагментов (булевские параметры) и физико-химические свойства структурных фрагментов или

молекул как целого (количественные параметры) [I]. К известным проблемам анализа молекулярных структур относятся следующие.

I. Проблема установления корреляций "структура-свойство", в том числе анализ и прогнозирование связи "структура-биологическая активность", компьютерная генерация молекулярных структур, расчеты физико-химических свойств молекул через структурные инкременты этих свойств.

2. Проблема классификации и представления молекулярных структур в химических базах данных. Важнейшие по практическому значению задачи первого направления связаны с решением проблемы конструирования химических соединений с заданными свойствами (для краткости будем называть такие соединения эффекторами). Большинство известных подходов к поиску новых эффекторов ориентировано на оптимизацию исследований в рядах химических соединений, поскольку описание молекул в них производится в терминах физико-химических параметров заместителей, связанных с неизменным остовом. Распространенный способ нахождения принципиально новых эффекторов основан на возможности использования информации о корреляциях между структурой и свойствами в одном ряду химических соединений на основании изучения этих корреляций в других рядах. Необходимыми условиями для этого являются описание молекул различных химических рядов и анализ связи "структура-свойство" на основе применения общих дескрипторов.

В настоящее время ясно, что основные затруднения при решении задач установления корреляций типа "структура-свойство" связаны именно с проблемой описания химических соединений (порождением дескрипторов) и с отбором из множества дескрипторов объясняющего подмножества-признаков (предикторов). Разумный выбор признаков требует использования содержательной информации о совокупности физико-химических процессов, происходящих при применении эффекторов, интуиции, метода проб и ошибок. Применяя статистические методы, можно определить несущественные ("шумовые") элементы описания, выбрать и упорядочить признаки в соответствии с некоторым критерием полезности, но нельзя найти новые (хорошие) признаки, не содержащиеся явно или неявно в исходных данных. В такой ситуации полезно иметь некоторую избыточность описания химической структуры эффекторов и физико-химических процессов, существенных для выполнения эффектором требуемых функций.

Например, механизм действия эффекторов-биологически активных соединений включает ряд стадий (введение/проникновение,транспорт/распределение, метаболизм, неспецифическое связывание с биологическими молекулами, выведение, связывание с рецептором и т.д.),являющихся по своей сути физико-химическими и описываемыми соответственно в терминах физико-химических параметров. В связи с этим весьма желательным является и описание структуры в терминах физико-химических параметров. Такое описание позволит решать задачу переноса связи "структура-активность" с одного химического класса на другие.

Элементы описания должны учитывать и тот факт, что для проявления биологической активности в большинстве случаев необходимо наличие двух или более центров взаимодействия (связывания, химического превращения) активной молекулы с рецептором. Нужно иметь в виду также, что для биологического действия могут оказаться существенными различные локальные характеристики молекулы биологически активного соединения (например, существенным свойством одного из заместителей может быть электронодонорная способность,а второго заместителя - гидрофобность или объем).

Таким образом, в проблеме связи молекулярного строения с биологической активностью естественно возникает необходимость рассмотрения дескрипторов, отражающих пространственные корреляции локальных физико-химических свойств на молекулярной структуре.

I. Характеристики распределения физических свойств на молекулярной структуре. Пусть молекулярная структура представлена в виде молекулярного графа - взвешенного неориентированного графа (определения см. в [2]) $G(R,W,P)$, где $R = \| r_{ij} \| = \{ r_{ij} / i,j = \overline{1,n} \}$ - матрица расстояний между вершинами графа, $W = \| w_{i\alpha} \| = \{ w_{i\alpha} / i = \overline{1,n};\ \alpha = \overline{1,N} \}$, $P = \| p_{k\eta} \| = \{ p_{k\eta} / k = \overline{1,m}; \eta = \overline{1,M} \}$ - матрицы весов вершин и ребер соответственно. Вторые индексы у величин $w_{i\alpha}$ и $p_{k\eta}$ нумеруют физические свойства, связанные соответственно с вершинами и ребрами (связями) молекулярного графа.

Ограничимся далее случаем наличия весов только у вершин графа и введем ряд характеристик распределений физических свойств на молекулярной структуре, являющихся некоторыми дискретными аналогами известных в статистике характеристик случайных процессов - корреляционных и кумулянтных функций [3] и метрических характеристик вершин графов [4].

Определим конфигурацию как подмножество из $t = 1,2,\ldots$ вершин графа, лежащих на некоторой простой (несамопересекающейся) цепи и задаваемых набором последовательных расстояний $D = \{d_1, d_2, \ldots, d_{t-1}\}$ между вершинами. Дискретные аналоги корреляционных и кумулянтных функций - спектры распределений физических свойств на молекулярной структуре, определим как результаты обобщенного усреднения (суммирования с весами) по множеству всех конфигураций с заданным набором D .

Ненормированные спектры введем как результаты простого суммирования по конфигурациям.

Ненормированные спектры первого порядка есть скаляры

$$S_\alpha = \sum_{i=1}^{n} w_{i\alpha} .$$

Совместный корреляционный спектр второго порядка физических величин α и β определим как вектор с компонентами

$$S_{\alpha\beta}(l) = \sum_{i,j=1}^{n} w_{i\alpha}\,\delta(r_{ij}-l)\,w_{j\beta}, \quad l = \overline{0,L}; \quad \alpha,\beta = \overline{1,N};$$

где $\delta(r-l)$ - символ Кронекера (дискретная дельта-функция):

$$\delta(r-l) = \begin{cases} 1 & \text{при } r = l, \\ 0 & \text{при } r \neq l. \end{cases}$$

Можно ввести и совместный корреляционный спектр второго порядка с учетом кратности - наличия нескольких (k_{ij}) цепей с одинаковым минимальным расстоянием между вершинами i и j :

$$S'_{\alpha\beta}(l) = \sum_{i,j=1}^{n} w_{i\alpha}\,k_{ij}\,\delta(r_{ij}-l)\,w_{j\beta} .$$

Аналогичным образом определяются и совместные корреляционные спектры высших порядков. В частности,

$$S_{\alpha\beta\gamma}(l_1,l_2) = \sum_{i,j,k=1}^{n} w_{i\alpha}\delta(r_{ij}-l_1)w_{j\beta}\delta(r_{jk}-l_2)w_{k\gamma} ,$$

$$S'_{\alpha\beta\gamma}(l_1,l_2) = \sum_{i,j,k=1}^{n} w_{i\alpha}k_{ij}\delta(r_{ij}-l_1)w_{j\beta}k_{jk}\delta(r_{jk}-l_2)w_{k\gamma} .$$

Для дальнейшего удобно ввести величины $w_{i0} = 1$, $i = \overline{1,n}$. Тогда

$$S_0 = \sum_{i=1}^{n} w_{i0} = \sum_{i=1}^{n} 1 = n \ ,$$

$$S_{00}(1) = \sum_{i,j=1}^{n} \delta(r_{ij}-1) \ ,$$

$$S'_{00}(1) = \sum_{i,j=1}^{n} k_{ij}\delta(r_{ij}-1) \ ,$$

$$S_{000}(l_1,l_2) = \sum_{i,j,k=1}^{n} \delta(r_{ij}-l_1)\delta(r_{jk}-l_2) \ ,$$

$$S'_{000}(l_1,l_2) = \sum_{i,j,k=1}^{n} k_{ij}\delta(r_{ij}-l_1)k_{jk}\delta(r_{jk}-l_2) \ .$$

Величины S_0, $S_{00}(1)$ и $S_{000}(l_1,l_2)$ есть числа конфигураций с одной, двумя (с расстоянием 1 между вершинами) и тремя (с расстояниями l_1 и l_2 между последовательными вершинами) вершинами соответственно.

Полные числа конфигураций

$$S_{00} = \sum_{l=0}^{L} S_{00}(1); \qquad S'_{00}(1) = \sum_{l=0}^{L} S'_{00}(1);$$

$$S_{000} = \sum_{l_1,l_2=0}^{L} S_{000}(l_1,l_2); \quad S'_{000} = \sum_{l_1,l_2=0}^{L} S'_{000}(l_1,l_2) \ .$$

Введенные величины нужны для использования в качестве нормирующих множителей при применении предложенного выше описания молекулярных структур к химическим соединениям с различным числом атомов с целью их классификации. Возможны, например, следующие нормирующие процедуры:

1) деление всех компонент величин S на число вершин в молекулярной структуре $S_0 = n$;

2) покомпонентное деление величин S на числа конфигураций $S_{00}(1)$ (или $S'_{00}(1)$) в случае спектров второго порядка и $S_{000}(l_1,l_2)$ (или $S'_{000}(l_1,l_2)$) в случае спектров третьего порядка;

3) деление всех компонент величин S на полное число конфигураций S_{00} (или S'_{00}) в случае спектров второго порядка и S_{000} (или S'_{000}) в случае спектров третьего порядка.

Можно рассмотреть и другие варианты нормировки - от отсутствия ее вообще (ненормированные спектры, введенные выше) до

наиболее общего случая усреднения в соответствии с некоторыми априорно заданными весами. Ограничимся случаем второй приведенной выше нормировки – нормировкой на единицу, когда при $w_{i\alpha} = 1$ для всех i и α все компоненты нормированных совместных корреляционных спектров равны I. Нормированные совместные корреляционные спектры

$$A_\alpha = S_\alpha / S_0 = S_\alpha / n ,$$

$$A_{\alpha\beta}(1) = S_{\alpha\beta}(1)/S_{00}(1),$$

$$A_{\alpha\beta\gamma}(1_1,1_2) = S_{\alpha\beta\gamma}(1_1,1_2)/S_{000}(1_1,1_2)$$

являются непосредственными аналогами совместных корреляционных функций случайных процессов [3].

Наиболее ясное описание статистических связей (корреляций) между случайными величинами (функциями) дают совместные кумулянтные функции [3]. Их прямыми аналогами в случае распределений физических свойств на молекулярных структурах являются совместные кумулянтные спектры

$$B_{\alpha\beta}(1) = A_{\alpha\beta}(1) - A_{\alpha 0}(1)A_{0\beta}(1),$$

$$B_{\alpha\beta\gamma}(1_1,1_2) = A_{\alpha\beta\gamma}(1_1,1_2) - A_{\alpha 00}(1_1,1_2)A_{0\beta\gamma}(1_1,1_2) -$$
$$- A_{0\beta 0}(1_1,1_2)A_{\alpha 0\gamma}(1_1,1_2) - A_{00\gamma}(1_1,1_2)A_{\alpha\beta 0}(1_1,1_2) +$$
$$+ 2A_{\alpha 00}(1_1,1_2)A_{0\beta 0}(1_1,1_2)A_{00\gamma}(1_1,1_2) .$$

Рассмотрим молекулярный граф G, представимый в виде множества связных подграфов R_k, $k = \overline{1,p}$, которые попарно не имеют общих вершин. В частности, один из подграфов может быть интерпретирован как остов молекулы, с которым мостами (ребрами, удаление любого из которых увеличивает число компонент связности) соединены подграфы – "заместители". Каждый из подграфов R_k, $k = \overline{1,p}$, можно охарактеризовать корреляционными и кумулянтными спектрами и рассматривать их наряду со спектрами S, A, B всей структуры как физико-химические дескрипторы при решении различных задач анализа молекулярных структур в химической информатике. В "предельном" случае подграф R_k может состоять из одной вершины. Для их описания, пользуясь определением слоев графа [4], введем спектры плотностей физических свойств по отношению к вершине i молекулярного графа соотношением

$$C_{i\alpha}(l) = \sum_j \delta(r_{ij}-l)w_{j\alpha}; \quad l = \overline{0,L}; \; i = \overline{1,n}; \; \alpha = \overline{1,N}.$$

Остановимся более подробно на ситуации, когда семейство молекулярных структур состоит из двух подграфов R и X , связанных мостом, причем подграф R в пределах семейства неизменен (случай химического ряда родственных молекул с остовом R и одним переменным заместителем X). Пусть мост инцидентен вершине $i \in R$,Тогда спектры $C_{i\alpha}(l)$ являются новыми дескрипторами, описывающими интегральное влияние переменных заместителей на молекулярную структуру.Используем вместо индекса обозначение RX и запишем: $C_{RX\alpha}(l) = C_{R(X)\alpha}(l) + C_{X(R)\alpha}(l)$, где

$$C_{R(X)\alpha}(l) = \sum_{j\in R} \delta(r_{ij}-l)w_{j\alpha},$$

$$C_{X(R)\alpha}(l) = \sum_{j\in X} \delta(r_{ij}-l)w_{j\alpha}. \qquad (1)$$

Величины $C_{R(X)\alpha}(l)$ и $C_{X(R)\alpha}(l)$, которые назовем спектрами плотностей физических свойств остова R и заместителя X соответственно, следует рассматривать как дескрипторы, характеризующие остов и заместитель с учетом их взаимодействия.

В физической химии обычно изучаются изменения свойств (в том числе реакционная способность, характеризуемая константами скоростей) в пределах рядов родственных молекул с одним или несколькими переменными заместителями. Одна из молекул ряда выбирается в качестве базисной, а отвечающие ей заместители X_0 называются стандартными (чаще всего это вершины $X_0 = H$). Для анализа такой ситуации введем следующие спектры:

$$\delta C_{RX\alpha}(l) = \delta C_{R(X)\alpha}(l) + \delta C_{X(R)\alpha}(l) , \qquad (2)$$

$$\delta C_{R(X)\alpha}(l) = C_{R(X)\alpha}(l) - C_{R(X_0)\alpha}(l), \qquad (3)$$

$$\delta C_{X(R)\alpha}(l) = C_{X(R)\alpha}(l) - C_{X_0(R)\alpha}(l). \qquad (4)$$

Для введения единой шкалы дескрипторов, характеризующих заместители, необходимо зафиксировать остов. В физической химии часто используются два набора констант заместителей, отвечающих определенным алифатическому и ароматическому остовам. Соответственно можно рассматривать и два набора дескрипторов (2)-(4) со стандартными подграфами-остовами, отвечающими алифатическому и ароматическому молекулярным фрагментам (например, n-пропильному и фенильному).

При использовании введенных дескрипторов для задач корреля - ционного анализа и классификации возможны следующие альтернативы. Применяемые в настоящее время константы заместителей (Гаммета, Тафта, Чартона и др.) могут быть корреляционно связаны со спектрами плотностей физических свойств (т.е. выражены через них) либо быть ортогональными (в статистическом смысле) к ним.

В первом случае введенные в данной работе дескрипторы позволяют исключить из математических моделей традиционные константы заместителей (по крайней мере, некоторые из них). А поскольку большинство констант заместителей в органической химии получено обработкой ограниченных рядов соединений (реакционных серий) и для них отсутствуют методы расчета для неизученных заместителей, замена их на расчетные дескрипторы типа введенных выше должна радикально расширить возможности теоретического анализа в физической органической химии.

В альтернативном случае спектры плотностей физических свойств следует рассматривать как дополнительные константы заместителей и использовать их в математических моделях (корреляционных уравнениях, алгоритмах классификации) совместно с традиционными константами заместителей.

Отправной точкой для развития рассмотренного подхода к анализу химических соединений, как распределени физических свойств на молекулярной структуре, явились работы [5,6], в которых предложен метод так называемого автокорреляционного дескриптора. Автокорреляционный дескриптор в терминах введенных выше определений есть ненормированный автокорреляционный спектр второго порядка $S_{\alpha\alpha}(l)$, $l = \overline{0,L}$.

2. Примеры. На рис.I приведены спектры распределений зарядов и гидрофобности для трех изомеров гексана. Инкременты зарядов на углеродных атомах и гидрофобности углеводородных групп взяты соответственно из [7] и [8]. Видна высокая чувствительность спектров к изомерным перестройкам молекулярной структуры.

Как простые примеры успешного использования введенных выше дескрипторов заместителей рассмотрим корреляции стерических констант Чартона ν и Тафта E_s^0 ([9,табл.Ш.3]) со спектрами плотностей $C_\alpha(l) \equiv C_{X(R)\alpha}(l)$, определенными соотношением (I), для алифатических заместителей C_3 - C_7 , где α - ван-дер-ваальсов радиус r атомов. Корреляции ищем в виде

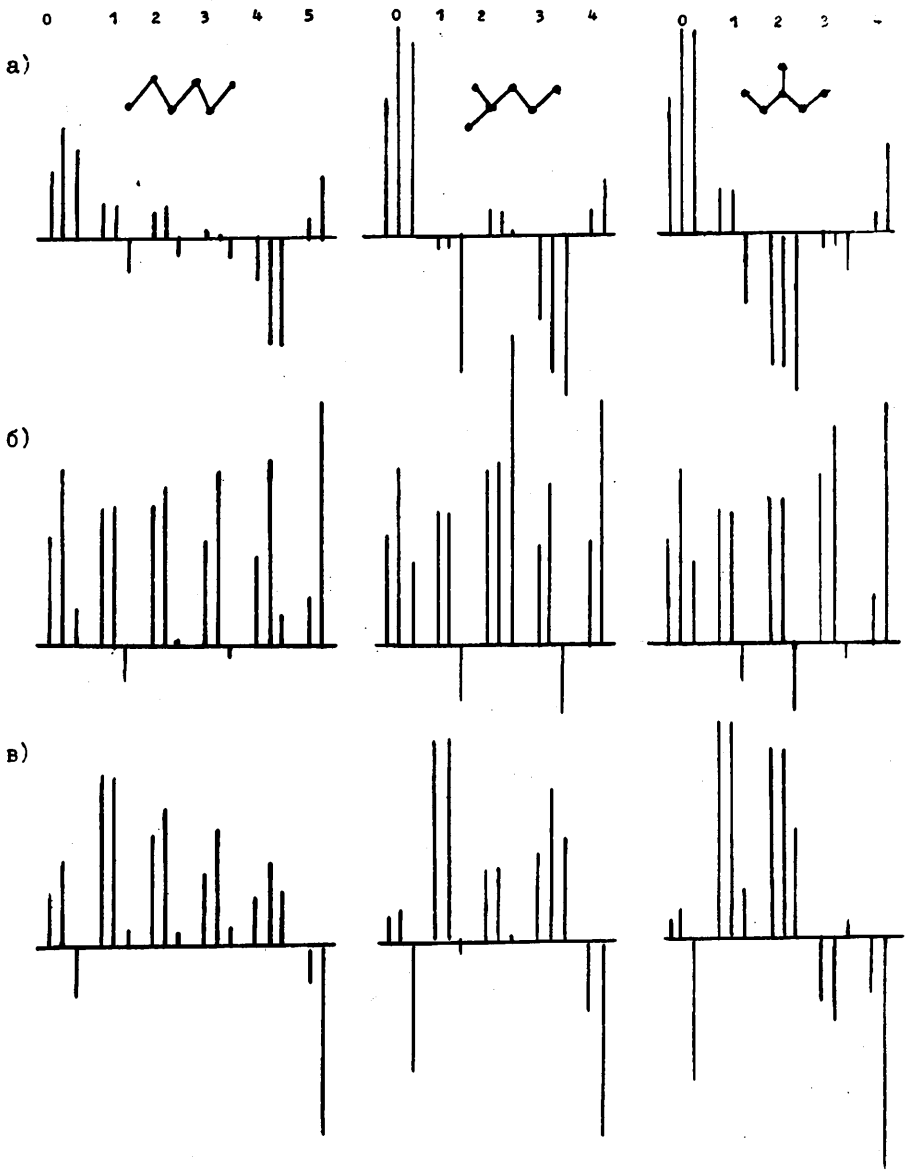

Рис. I. Спектры распределений зарядов q и гидрофобности f на молекулярных структурах изомеров гексана. Цифры вверху - расстояние l в числах связей. Компоненты при каждом значении l соответствуют величинам $S_{\alpha\beta}(l)$, $A_{\alpha\beta}(l)$, $B_{\alpha\beta}(l)$: а) $\alpha=\beta=q$; б) $\alpha=\beta=f$; в) $\alpha=q$, $\beta=f$

$$y = \sum_{l=1}^{L} a_l C_\alpha(l) + a_0, \quad y = v, E_s^0 . \qquad (5)$$

В силу соотношений $n_C(l) + n_H(l) = 3n_C(l-1)$ и $S_\alpha(1) = r_C = \text{const}$, где $n_C(l)$ и $n_H(l)$ – числа атомов С и Н в l-м слое, из соотношения (5) можно исключить вклады водородных атомов и объединить члены $a_0 + a_1 S_\alpha(1)$, в результате чего корреляция приобретает вид

$$y = b_0 + \sum_{l=2}^{L} b_l n_C(l).$$

Получены следующие модели:

$$v = -0,161 + 0,487 n_C(2) + 0,367 n_C(3),$$
$$n = 15, \; s = 0,107, \; F = 264, \; \rho = 0,988, \qquad (6)$$

$$-E_s^0 = -1,483 + 1,229 n_C(2) + 0,745 n_C(3),$$
$$n = 13, \; s = 0,121, \; F = 820, \; \rho = 0,997, \qquad (7)$$

где n – объем выборки, s – среднеквадратичная погрешность модели, F – значение критерия Фишера, ρ – коэффициент множественной корреляции.

Интересным следствием моделей (6)–(7) является заключение о независимости стерических констант v и E_s^0 от разветвленности заместителей на расстоянии, большем чем три связи от места замещения.

Заключение

Предложено новое описание химических соединений в терминах характеристик распределений физико-химических свойств на молекулярных структурах. Это описание может быть использовано при решении различных задач химической информатики и моделирования в физической химии с помощью имеющегося алгоритмического и программного обеспечения для многомерного анализа данных (см. обзоры [10, 11]).

Предложенный подход непосредственно обобщается и на трехмерные молекулярные структуры (путем дискретизации реального пространства и использования матрицы реальных межатомных расстояний вместо матрицы расстояний на графе).

Реализация метода требует создания систем атомных и групповых вкладов (весов вершин и ребер молекулярного графа или трехмер-

ной структуры) для различных физико-химических параметров, таких как поляризуемость, локальные атомные заряды, донорно-акцепторные факторы, логарифм коэффициента распределения (липофильность/гидрофобность), молекулярная рефракция и другие. Системы атомных и групповых вкладов (инкрементов) для многих физико-химических характеристик приведены в литературе (например, в [I2]).

Литература

I. BAWDEN D. Computerized Chemical Structure-Handling Techniques in Structure-Activity Studies and Molecular Prediction// J. Chem.Inf.Comput.Sci.-1983.-V.23.- P.14-22.

2. ГОРБАТОВ В.А. Основы дискретной математики. - М.: Высшая школа, I986. - 3I2 с.

3. МАЛАХОВ А.Н. Кумулянтный анализ случайных негауссовских процессов и их преобразований. - М.: Сов.радио, I978. - 376 с.

4. СКОРОБОГАТОВ В.А., ХВОРОСТОВ П.В. Анализ метрических свойств графов //Методы обнаружения закономерностей с помощью ЭВМ. - Новосибирск, I98I. - Вып. 9I: Вычислительные системы. - С.3-20.

5. BROTO P., MOREAU G., VANDYCKE C. Comparison of Logical or Three-Dimensional Molecular Structures by Means of Auto-Correlation Procedure//Computer Applications in Chemistry./ Eds.Heller S.R., Potenzone R.,Jr. - Amsterdam,1983.- P.263-284.

6. BROTO P., MOREAU G., VADYCKE C. Molecular Structures:Perception Autocorrelation Descriptor and SAR Studies. Autocorrelation Descriptor//Eur.J.Med.Chem.- 1984.- V.19,N 1.-P.66-70.

7. МАЛИЧЕНКО Б.Ф. Молекулярные диаграммы органических соединений. - Киев: Наукова думка, I982. - 228 с.

8. REKKER R.F., KORT de,H.M. The Hydrophobic Fragmental Constant: an Extension to a 1000 Data Point Set// Eur.J.Med.Chem., 1979.-Vol.14,N 6.- P.479-488.

9. ПАЛЬМ В.А. Основы количественной теории органических реакций. - Л.: Химия, I977. - 360 с.

I0. ВАЛУЕВА Л.Н., ЗАЦЕПИН Б.М., ПРОМОНЕНКОВ В.К. Применение математических методов для анализа связи молекулярная структура - пестицидная активность (часть I). - М.: НИИТЭХИМ, I985.- 56 с. -(Хим. пром-сть и пром-сть по произв. мин.удобр. Химические средства защиты растений: Обзор. инф.).

II. Их же. Применение математических методов для анализа связи молекулярная структура - пестицидная активность (часть 2).-М.: НИИТЭХИМ, I986. - 54 с. -(Хим.пром-сть и пром-сть по произ. мин. удобр. Химические средства защиты растений: Обзор. инф.).

I2. РИД Р., ПРАУСНИЦ Дж., ШЕРВУД Т. Свойства газов и жидкостей. - Л.: Химия, I982. - 592 с.

Поступила в ред.-изд.отд.
I3 ноября I986 года